Finite Mixture

Thursday, March 19, 2020 6:44 PM

Finite Mixture

- Smoothing parameter, h, free method
- Reduction of Computational burden
- Conditions may not be applicable for many applications

Univariate Finite Mixture

$$f(x) = \sum_{i=1}^{c} p_i \cdot g(x; \underline{\theta}_i), \quad c << n$$

p_i: weight, $\sum_{i=1}^{c} p_i = 1$, $p_i > 0$

$g(x; \underline{\theta}_i)$: component density with parameter $\underline{\theta}_i$ at x

; weighted sum of the density $g(x; \underline{\theta}_i)$

- The component density can be either continuous or discrete.
  - continuous: usually normal density
- # of component density, c, need to be known in advance
- p_i and $\underline{\theta}_i$, $i = 1, \cdots, c$ are to be estimated.

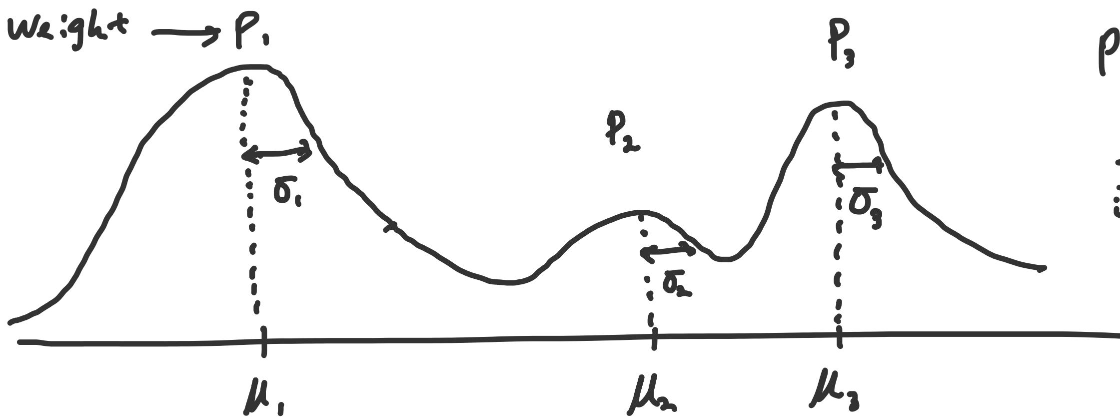
For normal component density,

$$\underbrace{\hat{f}_{FM}}_{\text{Finite Mixture}}(x) = \sum_{i=1}^{c} \hat{p}_i \cdot \underbrace{\phi}_{\text{normal density function}}(x; \hat{\mu}_i, \hat{\sigma}_i^2)$$

- The observation x comes from one of these normal distribution but we do not observe to which component it belongs.
- Parameters to be estimated ; $p_1, p_2, \cdots, p_{c-1}, \mu_1, \mu_2, \cdots, \mu_c$
  $\sigma_1^2, \sigma_2^2, \cdots, \sigma_c^2$

  Total # of parameters $= 3c-1$

Ex



$p_1 > p_3 > p_2$

$\sum_{i=1}^{3} p_i = 1$

Example 9.8 Univariate Finite Mixture model with 3 normal Components

$$f(x) = \underset{p_1}{0.3} \times \phi(x; \underset{\mu_1}{-3}, \underset{\sigma_1}{1}) + \underset{p_2}{0.3} \times \phi(x; \underset{\mu_2}{0}, \underset{\sigma_2}{1}) + \underset{p_3}{0.4} \times \phi(x; \underset{\mu_3}{2}, \underset{\sigma_3}{0.5})$$

See the sample code.

- Once a sample data set is given, we need to identify the number of components, c, and the component density.
- EM (Expectation Maximization) algorithm can be used

to estimate p's, μ's and σ's.

Finite Mixture vs Kernel density estimation

$$\hat{f}_{FM}(x) = \sum_{i=1}^{c} \underbrace{\hat{p}_i}_{\text{weight}} \cdot \underbrace{g(x; \hat{\underline{\theta}}_i)}_{\text{component density}} \quad \text{vs} \quad \hat{f}_{ker}(x) = \underbrace{\frac{1}{n}}_{\text{equal weight}} \sum_{i=1}^{n} \underbrace{K_h(x - X_i)}_{\text{Kernel density as a function of } h}$$

- Kernel est is a special case of FM where $c = n$

Parameter Estimation

EM algorithm (Dempster, Laird and Rubin '77)

E → Expectation, M → Maximization

Optimizing likelihood function with missing data

[csfinmix]

Iterative Procedure (Univariate)

Normal component density case

Step 1 : Determine c (exploratory data analysis can be used)

Step 2 : Determine initial guess for $\hat{p}_i$'s, $\hat{\mu}_i$'s and $\hat{\sigma}_i$'s

Step 3 : Calculate the estimated posterior prob. that x_j (data)

$j = 1, \ldots, n$ belongs to

$$\hat{\tau}_{ij} = \frac{\hat{p}_i \cdot \phi(x_j : \hat{\mu}_j, \hat{\sigma}_j^2)}{\hat{f}(x_j)}$$

$$\hat{f}(x_j) = \sum_{k=1}^{c} \hat{p}_k \cdot \phi(x_j : \hat{\mu}_k, \hat{\sigma}_k^2)$$

Prob. that x_j belongs to $\phi(\hat{\mu}_i, \sigma_i^2)$, i^{th} component.

Step 4 : Update

$$\hat{p}_i = \frac{1}{n} \sum_{j=1}^{n} \hat{\tau}_{ij}$$

$$\hat{\mu}_i = \frac{1}{n} \sum_{j=1}^{n} \frac{\hat{\tau}_{ij} x_j}{\hat{p}_i} \quad : \text{weighted average}$$

$$\hat{\sigma}_i^2 = \frac{1}{n} \sum_{j=1}^{n} \frac{\hat{\tau}_{ij} (x_j - \mu_i)^2}{\hat{p}_i}$$

Step 5 : Repeat step 3 and 4 until converge

<See Example 9.13>

Generating r.v. from the finite Mixture model

Step 1 : Get the FM model : $p_i , g_i(x ; \underline{\theta}_i)$, $i=1,\dots,c$

Step 2 : Generate n r.v. from $U(0,1)$; $u_1, u_2, \dots, u_n$

Step 3 : Count # of u's in $[0, p_1) \Rightarrow n_1$

" " in $[p_1, p_1+p_2) \Rightarrow n_2$

$\vdots$ $\vdots$ $\vdots$

Count # of u's in $[p_1+\dots+p_{c-1}, 1] \Rightarrow n_c$

Step 4 : Generate n_1 r.v. , $X_1, \dots, X_{n_1}$ from g_1

n_2 r.v , $X_{n_1+1}, \dots, X_{n_1+n_2}$ " g_2

$\vdots$ $\vdots$

n_c " " " g_c

< See example 9.13 >

$\boxed{\text{In-class assignment}}$

1. Write a procedure to generate r.v. from normal Kernel density estimation (similar to the procedure for FM)

2. (a) Create artificial 3-term mixture data with $n = 1500$ of which

$n_1 = 200$ from $N(5, 3^2)$

$n_2 = 800$ from $N(10, 1.5^2)$
$n_3 = 500$ from $N(15, 2^2)$

This is your raw data.

(b) Find the finite mixture model using EM
[csfinmix]

Use the initial $p = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$
$\mu = [4, 12, 16]$
$\sigma = [2, 1, 2]$

(c) Now, generate $n = 1500$ r.v. from the finite mixture model.

(d) Draw density histogram of the raw data in (a) and r.s. in (c). Compare.

3. Using the same raw data in #2, estimate the density using normal Kernel density estimation method. Superimpose the density estimation on the histogram of the raw data.

Repeat (c) and (d) in #2.